言語研究のための外国語テキストデータ検索：ウィンドウズ環境の場合

出口厚実

Procesamiento de textos para la investigación lingüística en el entorno Windows

Atsumi DEGUCHI

El sistema Windows Japonés de Microsoft sigue ganando cada vez más usuarios en este país desde que se lanzó su versión 3.1 en el mercado local a mediados del año 1993.

Este entorno operativo con el interface gráfico de usuario (GUI), aunque ayuda a hacer más eficaces y fáciles algunos tipos de gestiones del trabajo cotidiano, implica unos problemas serios a los que se dedican a los procesamientos de textos en idiomas extranjeros para la investigación lingüística. En Windows Japonés no se puede representar ni teclear caracteres con diacrítico indispensables para la ortografía española.

Otro punto frágil del nuevo sistema es la insuficiencia y escasez de utilidades relativas a la manipulación de textos ASCII puros (o sea ficheros en formato de texto estándar propio del sistema).

En vista de esta situación el autor ha desarrollado los tres programas siguientes que funcionan bajo Windows con miras a la utilización de los investigadores y estudiantes de la lingüística enpañola:

wwfi.exe :localiza palabras en los ficheros de texto ANSI y facilita la edición de los datos obtenidos

wwsch.exe :ofrece la posibilidad de editar textos en español con distintos colores de fondo y tipos de letra de pantalla

dfind.exe :adopta el método Drag-and-Drop para arrancar el programa, permitiendo presentar y modificar los resultados de la búsqueda por medio del editor de textos especificado por el usuario

0．序

テキスト検索は言語研究におけるコンピュータ利用の原点であり、また実際に最も多く活用されている基本操作であることに異論はないであろう。

特定のプロジェクトのために定格化された言語コーパスは無論のこと、不統一な雑多なテキストの寄せ集めであっても、およそ言語の自然な断片であれば、規模の大小を問わずあらゆるテキスト・データは、その言語を調査研究するものにとって有用である。用例の採取、語彙・文法形式の分布、共起文脈の探査や検証など外国語研究の様々な局面で、常時、問題となる現象の実証を要求されているからである。実例に基づく裏付けも、1、2の例だけでなくできるだけ多数の中から好例を捜したいという欲求に駆られることは度々経験する。電子化されたデータにアクセスすることができ、かつそれを手軽に扱える用具が存在するならば、誰しもテキスト検索を研究の補助に役立ててみたいと思う。

コンピュータ言語学の専門でない一般の外国語研究者、学習者がこの便利な手段を十分に活用できるか否かに3つの条件が係っている：

第1はデータの存在とその使用許可の問題である。テキスト検索の利便に与りたいがテキストデータそのものが利用できない、あるいはその所在が不明であるとか、あるにはあるがデータが開放されていないというケースも多い(1)。また、高価すぎる利用料金が妨げとなる場合もある。

第2は必要な器材が準備されているかである。パソコンの低価格化の伴う急速な普及により、現在、普及型と称される水準のハードウェアでもかなり大規模なテキストを処理するのに十分な能力を備えるようになっている。

最後に効率的な検索ソフトが手に入るかどうかも重要である。上述の2条件は完全に満たされていながら、この点でテキスト検索の利点を享受し切らないでいる人々は案外多いようである。現在のところ、個人レベルで言語データベースを処理するためのソフトウェアが十分揃っているとは言いがたい(2)。それどころか、各種のツール類をどこで、どのようにして見つけたらよいのかというような、門前情報すら欠けている状況だろう。素材と道具立てが揃っていても、それを実際に導入して利用するかどうかの決断には、操作環境がわかり易く、特別な訓練を必要とせず、呪文めいたコマンドをいくつも覚えなくても、たやすく目的が達せられるかどうかも重要である。

パソコン利用のデータ検索に効用を認め、実践を検討している多くの人々を躊躇させたり断念させた一因は、根本的な“利用可能なコーパス、テキストデータの乏しさ”があるにしても、ハード及びソフトを使いこなすことにより得られる成果とそこにいたるまでの手続きと準備の煩雑さに対するバランス感覚が先立つためであろうと思われる。

昨年来、我が国でもパソコンの動作環境としての MS-Windows が急速に普及してきた。ウィンドウズがパソコンの初心者やテキスト処理の入門者にとって、より好適な土台になり得るならば、そこでのテキストデータ検索も、上述のソフト面における障壁を軽減する効果をもたらすか

もしれないという期待がある。本稿はこのシステムが外国語（特に断らない限りこの紙面ではスペイン語を取り上げるが、ラテン文字＋特殊字母を用いるヨーロッパ各国語に共通する問題である）テキストデータを処理する上で、どのような問題を含んでいるか、それらはどのようにして改善または解決可能か、焦点をデータ検索に絞りながら考察してみたい。

１．ＧＵＩは不用か

コンピュータのオペレーティングシステムとして Graphical User Interface が徐々に一般化してきている。OS全体に及ばなくても、その一部に食い込んだ、あるいは包括的なアプリケーションを管理する基幹アプリケーションとしての GUI は、遅れ馳せながら我が国のパソコン界においてもやがて支配的になるのではないかと予想されている。一方、ウィンドウズのような GUI はテキスト処理に不用であるという意見がある。確かに、テキストデータ処理は視覚的な図形に補完されて益する部分が少ないのは事実である。また、GUI 環境を採用することにより、CPU の能力にオーバーヘッドがかかり、全般に処理スピードが低下するといわれる。テキスト処理用のものであれ、ソフトウエアの開発はグラフィックユーザインターフェースの無い通常のテキストベースのそれに比べて複雑化し時間資源を浪費する。さらに、Windows 用のソフトを実行するための最低限のハード条件は、従来のテキストベースのそれに比べて数段と厳しくなるという欠点もある（ハードディスクが必須；必須メモリー量も大きい；システムが要求するメモリー量も、個別アプリが消費するメモリー量も大きい）。

これらのマイナスの側面にもかかわらず、筆者はテキスト処理環境でのGUI 不必要論には賛成しない。パソコンでのテキストプロセシングがすべて早急に GUI へ移行しなければならないとは考えないが、早晩、全てのソフトウエア実行環境が GUI 化する趨勢の中で、この種の作業は例外的に DOS 上で行われ続けなければならないという理由を見出すことができないからである。現在、MS-DOSのような非 GUI のシステムと Windows との関係に代表されるような両体系が競合共存する段階でも、後者におけるテキスト処理が、もし、より快適な使い易い環境を提供するならば、積極的にその存在意義を認めてもよいのではないかという意見である。

GUI の基本的特色を弁護したり、その長所短所を理論的に考察する場ではないので、テキストデータを扱う立場から、ただ１つ次のようなメリットが確実ならば、GUI のシステム下でデータ検索などを出来るようにするべきだという結論に達する：

「MOS-DOS あるいは UNIX などのコマンド処理を中心とするOSよりも、視覚性を高めオブジェクト化を進める GUI 環境の方が、パソコンに不慣れな多数のユーザにとって、操作法を習得しやすく、また運用しやすい」

専門化した特殊な目的のデータ処理、大規模テキストの集計、その他の特化された複雑な定型処理のために非GUI 環境の方が効率的であれば、もちろんそのシステムを継続すればよい。ス

ピード至上主義にこだわるならば、恐らく非 GUI のテキスト処理を手放したくないだろうし、それも１つの選択である。

上に指摘したように、言語研究・教育に従事しながら、テキストのデータ処理にパソコンを利用しない人が意外に多い原因の１つが、DOS 環境のソフトウエアの機能不足ではなく、その操作の馴染みにくさにあるとすれば、この壁を取り除けば、さらに多くの人々がこれを実践するかもしれない。

データ検索やその結果の処理技術が何か秘術的な特技として、それ自身が新しい方法論として価値あるかのごとく秘匿したり、閉鎖的な世界に閉じこめることを当然視するような風潮が見られるのは残念なことである。

もし、従来のようなおびただしい文献の多読、長期間の精読から得られるのと同等の、またはそれ以上の成果をもたらす効率的な探索法があるならば、それを他の人々にも使用できるように紹介し、アイデアや情報を自由に交換・交流させるようなオープンな協力体制こそ必要である。初めから完備された快適な利用環境を望むのは楽観的すぎるとしても、試習的跛行と試行錯誤の積み重ねがより安全で便利な状況を産み出し、利用者の増大を呼び起こし、またそれがソフトウエアの改良へとフィードバックされることになるはずである。

GUI をサポートする Hard＋Soft の環境は、現在、いくつか提供されている。それらの中ではローエンドに位置し、従って最も個人で利用しやすく、また普及しているのは Mac OS と MS-Windows である。これまでの DOS の普及度から見て、MS-Windows の利用者及び未来の利用者が多数を占めるだろうと予想されるほか、現在のところ、筆者自身がわずかながら使用経験を持つ唯一のプラットフォームであるという理由で、以下の考察はすべて、Microsoft Windows ver 3.1を対象にしている。

GUI のシステムの比較検討から Windows がどの程度優れているか、あるいはどのような問題点があるかなどの一般的議論を提起したり、論評を加えるのは本論の射程を越えており、以下でのいくつかの考察は、Windows 以前の MS-DOS 環境との違いが実務上どのように問題となるかという視点からなされている点を予めお断りしなければならない。また、パソコン用市販ソフトや Freeware, Shareware の新製品は極めて早いテンポで開発され、市場に投入されるため、この稿が印刷される時点で、便利なツールが流通し始めているかもしれない。外国語テキスト処理に適したソフトに関しても筆者が知り得ている情報はごく狭い範囲のものであるため、ここで言及されていない優れた utility の類が存在し活用されている可能性も十分あり得る。

２．MS-Windows でのテキストデータ処理

GUI に基づいたシステムがその基本的な使い易さのゆえに、コンピュータの初心者や、他のメインな業務の合間に時折テキスト検索を行う非常習的エンドユーザにとって、有益なものと仮定することと、それが実現されるような環境がすでに整備されているかどうかは別の問題である。

この節では、後者の実情を確かめるとともに、Windows 上で外国語を扱う実作業に際して生じる様々な課題を概観し、その不備を幾分でも改善する具体策を探るのを目的とする。

2．1　日本語 Windows と外国語処理

国内のパソコン用に Windows がリリースされ、多くのユーザを獲得し始めたのは日本語 Windows ver 3.0以降である。AT互換機を対象とした PC-DOS／V あるいは MS-DOS／V 用の Windows にせよ、NEC9800（及びその互換機）用の Windows にせよ、「日本語」Windows として市販されているシステム ver 3.1 では、「日本語」MS-DOS がこれまでそうであったように、欧文特殊文字にほとんど対応していない。フォント切り替えによる他言語表示に対する柔軟性という GUI が本来持っている特長が依然として生かされようとしないのをみるとき、「外国語」＝「英語」という国民的ワンパタンはそうたやすく払拭されるものではないと改めて感じざるを得ない。すなわち、出荷されたままの状態では、例えば、スペイン語の特殊文字を表示することも入力することもできない(3)。

2．2　Windows の標準テキスト

MS-Windows でテキストデータを作成したり、テキストに何らかの処理を行う場合に、一番注意しなければならないのはテキストのフォーマットである。Windows が MS-DOS 上で動作する１つの application であるという基盤にもかかわらず、日本語以外の欧文テキストの文字コード割り当てが DOS テキストと Windows 用テキストで異なるためである。厳密に言えば日英語でも通常の字母、句読点以外の各種記号（図形文字）に関してもこのような食い違いが生じ得るのであるが、特にスペイン語などのように、特殊文字と呼ばれるアクセント付き母音などの内部処理コードが DOS／Windows 間で一致しないということは、これまでに蓄積されて来たデータがそのままでは流用できないという厄介な事態を引き起こす。

実際、日本語 MS-Windows は日本語版の DOS とのデータ互換性しか配慮されていないようで、多言語対応のコードページ850や英語系 DOS に基づいて作られたテキストデータが持ち込まれ利用されるケースを無視している。ウィンドウズに付属する文書作成プログラムの『ライト』及び『メモ帳』のいずれにおいても、PC-DOS／MS-DOS の英語モード上で作成された非英欧文データ中の ASCII CODE 128以上の文字は正しく表示されず、いわゆる文字化けを起こす。『ライト』ではいく種類かの欧文フォントを表示用に選択できるにもかかわらず、これらの文字をキーボードから入力しようとするとbeepで入力不能であると注意を受ける。フォント変更がサポートされない『メモ帳』では当然表示も入力もまったく不可能である。

AT互換機の DOS／V システム上では日本語モードと英語モードが簡単に切り替え可能であり、欧文テキストの処理に重宝されているが、DOS／V 用の Windows システムでこの機能が削られたのは、テキストの国際化、互換促進の面から見るとむしろ一歩後退であり非常に残念な

ことである。

ワープロ以外の MS-Windows のアプリケーションで非英欧文を画面に表示したり、データとして入力する方法として、現在思いつく方法は次の４つである。

（1）

1．日本語 Windows／英語 Windows の両システムを用意し必要に応じて切り替える

2．日英両 Windows ソフトを実行可能な日本語 OS／2 を導入する

3．日本語 Windows 用アプリケーションで非英欧文処理に対応するものを用いる

4．日本語 Windows で非英欧文処理に対応する独自のソフトを開発する

それでなくても肥大化した日本語 Windows システムに加えて、disk space を大きく占有する英語 Windows を別途に入手して、両方をインストールしておかなければならないという１．の解決策はあまりにも無駄が大きいようだ。日本語環境と外国語環境を同時に両立させることができないのも不便である。

1993年末にリリースされた OS／2 ver 2. 1J は基本システム内に Win OS／2 という Windows 互換セッションを持つ。しかも日本語 Win と英語 Win に別々に対応できる仕組みになっている。もしこれが手軽に実現できるのならば、非英欧文に加え通常の日本語処理も可能になり理想的な環境に近いことになる。ただし、OS／2自身が、Windows 以上に巨大なオペレーティングシステムと化して、誰でもがローエンドの入門機に導入することができるものではないという制約を覚悟せねばならない。もう１つの不安材料は、英語 Win OS／2 が、英語版 Windows と同等な多国語を本当に支援しているかという懸念である。乏しい使用経験のため、断定的な結論は避けたいが、英語 Windows では何ら問題のない、一部の欧文特殊文字（例えばスペイン語で必須の ñ, Ñ, ¡ など）の入力が現 version では拒絶され、英語モード対応とうたわれているのは、実は狭義の米国仕様を指しているのではないかと思われるふしがあるからである。

すでに手持ちのシステム単体でスペイン語などの外国語テキストをウィンドウズで処理する近道は、結局３，４に限られることになる。

既に市販されていたり、フリーウエアとして流通中のもので、上の３．に合致するものを見つけ出し、それらを組み合わせてうまく適合させることが可能ならば、なるべくこの方針を進めるのが得策である。

我々が要望するような非英外国語のためのテキスト処理ツールが、日本語 Windows で動作することは原理的に不可能ではないので、将来、積極的に開発されて公表されれば、最も望ましいのであるが、上記４．の打開策は現状では皆無に近い。MS-DOS に比べてプログラミングが煩雑であるのが障害となっていると思われるが、多言語処理のためのシステム改変法や、関連の Windows 関数の日英版での違いなどに関する技術資料が一般のユーザに入手しにくいのが災いしている面も見逃せない。

基本ソフトの提供者がシステムを設計する際に、非英欧文に関して、ほんのわずか配慮を加えたり、オプションで変更可能な仕様にしておけば容易に解決できた事柄が、見過ごされたり過小に評価されたりしたために、現在の不便な状況を作り出しているのではないかと悔やまれる。

2．3　検索のためのツール

テキスト中の文字列を検索する用具として DOS system には標準で find という外部コマンドが含まれている。拙稿（1990b）でも触れたが、これは我々が自然言語テキストの中の単語や語彙データを探索する用途としては、あまりにも貧弱で、それに代わる簡便な常備用のツールを新たに作成する必要性が生じた。ところが、MS-Windows システム自身には、この種の素朴な検索プログラムさえも添付されていない。つまり、ウィンドウズの環境で、もしテキストの文字列検索をしたければ、全面 DOS モードに戻るかDOS 窓を開いて、その中で MS-DOS の find を実行すればよい、という姿勢であろう。『テキスト処理に関しては Windows は不得手なので、DOS でやってください』とでも解釈できる扱いである。

前節で指摘した課題と、データ検索（ここではスペイン語語彙の検索が取り敢えず急務である）の便宜を加えて、最低限度下記の条件をさしあたって解決しなければならなかった。

（2）

日本語 Windows のテキスト処理において

1．非英欧文の特殊字が表示出来るようにする

2．非英欧文の特殊文字を含む文字列を検索可能にする

3．非英欧文の特殊字を Dead Key 方式で入力しやすくする

4．MS-DOS で使用されている諸々のデータ形式との互換を保障する

日本語 Windows 上で動作するエディタの中には上記１，２をサポートしているものもある（勿論、On memory での検索であるが）。一方、海外製ソフトの中にも、欧文系フォントを利用しても文字化けを妨げず、また Alt＋数字での文字入力を受け付けないものも見られる。ウィンドウズ用のファイル文字列検索ソフトがいくつか流通しており、入手可能なものを調べてみたが、欧文特殊字を表示入力できるものはなく、また表示フォントを変更できるものも見当たらなかった(4)。

また、これらは例外なく一般的な DOS 版検索ツールを意識して作成されているので、拙稿（1990b）で問題にしたような様々な不便さをそのまま引き継いでいる。とりわけ、致命的なのは３つの不対応がやはり残されている点である。

（3）

a．ハイフンで分離された「語」に対応していない

b．発見箇所のページ数を知らせない

c．単語を含む「文」を切り出してくれない

GUI を利用したマルチウィンドウ処理など、DOS にない新たな利便性は加えられたが、外国語テキストデータを日常処理したいものにとって、上の諸課題が解決されない限り、ただ Windows で文書が検索できるというだけでは、新しい環境に移行しようという気が起こらないかも知れない。

3．Windows 用検索プログラムの自作

以上のような、Windows システムと外国語との非親和性を、素人の手でいくらかでも取り除きスペイン語のテキストデータを処理しやすくできないかと検討した結果が、次節3．1－4に紹介する4種類の私作ソフトである。これらは、もともと手作りによる個人使用を出発点としていたため、商業ソフトのような多機能・汎用性を備えていない。また、現在も改良中のため、追加すべき機能で実装されていないものがある他、実現できないで残された課題も多い。これらの点に関し詳しい技術情報をお持ちの方や、すでに解決法を見出された方の御教示が得られれば幸いである。

3．1　WWF.EXE (Win Word Finder ver. 0.9)

『日本語 Windows 対応スペイン語単語検索プログラム』

前節（3）の要件に加えて、従来の find, grep 等による検索作業で最も不便を感じていたのは、消え去った検索結果画面を再度見直したり、確認してから保存するか、しないかを取捨するという選択ができない点であった。dos コマンドの more はページバックが効かないし、テキストバッファをスクロール可能にするユーティリティは幾種か流通するが、残念ながら外国語特殊文字には対応しない。もし可能であっても、編集するにはエディタに結果を取り込む必要があり面倒である。検索から得られた結果が直接 Windows 上に提示できるようになれば非常に好都合である。繰り返して出力を見られる他、テキストそのものを直ちに削除・追加など修正できるためである。

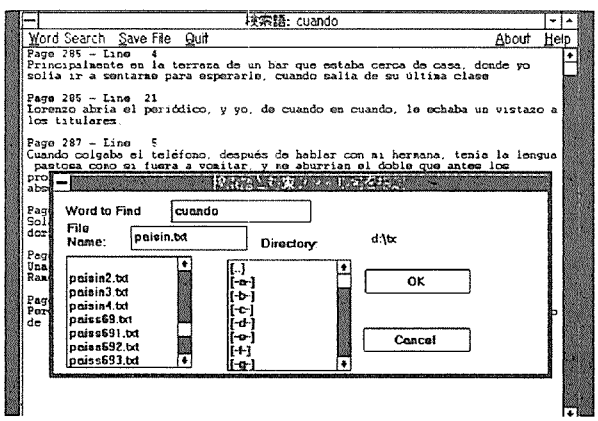
テキスト編集でのスペイン語表示は、システムフォント以外の欧文系フォントや固定 ansi フォントに交換してやれば、日本語ウィンドウズでも可能になることは、試行錯誤の後に判明した。検索後のテキストウィンドウでは、従って、スペイン語表示が実現するだけでなく、さらに edit control をサブクラス化して、dead key による特殊字の入力に対応させることも可能になった。ところが、一般ダイアログボックスや Windows 関数に組み込まれたユーザ入力窓では、キーボード入力の際に、カタカナ領域のコードを除いて Ascii code No.128 以高を受け付けないという現象が起こる。すなわちスペイン語で必要な特殊文字・記号のうち、Á, Í, Ñ, Ó, ¿, ª, º,

《, 》を除いて、常用のアクセント付き文字がビープ音と共に入力が阻止されてしまうのである。恐らく何らかの回避法があるのではないかと思われるが、全キー入力を監視しながらテキスト表示を自前処理して入力文字を獲得するルーチンを設けるのは手間がかかるので、未対応のままに残している。

簡単な抜け道としては検索語の入力の際には、STX 形式（あるいは特殊記号を用いないその他の方式）の表記を用い、内部処理データを ansi 形式に変換する策があるが、1つのソフトで両様の表示法を使い分けるのは混乱を招く恐れがあると判断した。その結果、取り敢えずSTX換字方式のテキストを検索でき、短いスペイン語テキストを編集したりセーブするための、妥協的な用途に限定することとした。

もう1つの制限は、かなり厳しいものであるが、Windows システムの Edit Control を利用する代償として、テキスト容量の上限が32Kバイトくらいに限られる点である。500 Kbyte ほどのファイルで高頻度語を検索すると、結果出力はこの限度を越すことは十分あり得る。簡易なテキスト処理で済ます限り、このリミットを克服することはかなり難しいので、後述3．4のような別法を編み出さざるを得なかった。

ただし、ふつうのサイズの複数テキストを順々に探索するのにはこの wwf.exe の方が、より高速で快適なのは確かであろう。使い勝手の良さと引き換えに、検索速度は DOS 環境よりかなり落ちるだろうという悲観的予想を裏切り、200-300 Kbyte の中編小説だと、ほとんど一瞬に検索結果画面が表示されるので、スピードに関してDOS 版の wfind.com などと伍する程で十分に実用になると感じた[5]。



[図1] wwf. exeでの検索結果画面と検索語入力・ファイル指定ダイアログ

この検索ソフトの具体的な内容は、wwf に添付の document を読めば明らかになると思うので、その全文を次に記載して紹介しておく。なお、これは起動後 help メニューの選択によって現れる説明文書と同一のものである。

>>

日本語 MS Windows 対応
スペイン語単語検索プログラム “WWFIND”
wwf.exe　　ver. 0.9　　by Atsumi Deguchi

単語検索をウィンドウズ上で行い、その検索結果の編集や、他の Win アプリケーションとの間でカット&ペーストを可能にする軽便なプログラムです。

◆◆特徴◆◆

1．MS-DOS の標準テキスト形式（スペイン語文）の文の中の特定の単語を検索し、それを含む「文」を切り出して、ウィンドウ上に出力します。
2．検索結果は、文字の削除・挿入など編集ができるほか、Windows ソフト間で切取り、張り付けが可能で、ファイルとして保存することもできます。
3．検索対象ファイルの選択は、他の Win アプリと同様、マウス操作により、ドライブ・サブディレクトリ・ファイルのダイアログ・リストから指定します。
4．NEC98（及び Epson の互換機）などの日本語 MS-Windows から起動した wwf の編集ウィンドウでもスペイン語特殊文字を表示し、dead key により入力できるので簡易エディタとしても利用できます。

◆◆制限と注意◆◆

1．検索対象テキストはSTX形式を想定していますが、他のフォーマットの MS-DOS、Windows ANSI テキストも扱うことができます。ただし、「文字列照合ルーチン」は拙作の wf.exe と同じ処理をするので、文頭の大文字→小文字変換が正常に機能しないケースが考えられます。STXでは@（逆疑問符）と｜（逆感嘆符）直後の文字も文頭と見なしていますが、この判断条件がテキスト書式によって変化するためです。
2．検索語の入力ボックスでは殆どの欧州語特殊字の入力が出来ません。ただしそのコードが半角カタカナに対応するならば検索文字として有効で、IBM 拡張 ASCII を使用するテキストや Windows ANSI テキストも正常に検索されます。
3．プログラム中を通じて全角文字を入力することはできません。

◆◆プログラムの起動◆◆

＜起動＞

wwfind の実行に必須のファイルは wwf.exe だけですが、wwf09.hlp を同じ directory にコピーしておけば、この説明文書をHELPメニューで呼び出すことができます。起動の仕方は次の 3 通

りあります：

1．予めいずれかのプログラムグループに登録しておいたwwfindのアイコンをダブルクリックする

2．プログラムマネージャの「ファイル名を指定して実行」を選び、wwf.exe のあるパスとファイル名を指定する

3．ファイルマネージャを起動し、wwf.exe をクリックする

＜検索語、検索ファイルの指定＞

プログラムが開始するとすぐに、検索単語の入力と対象ファイル名を求めるダイアログ・ウィンドウが開きます。ここでマウスを検索語入力ボックスへ移動してクリックすれば、文字入力モードとなりますので、単語（20字まで）を原則として小文字で入力して下さい。また対象ファイルがカレントディレクトリのリストの中にあれば、それをダブルクリックすると、直ちに検索を開始します。Drive, Sub-directory を変更するには右側のディレクトリボックスを順にクリックして左側のファイルリストにファイル名が現れるようにし、それをクリックします。

指定を終えれば右のＯＫボタンを押します。もしここで、Cancel を押すと、前回の検索結果表示ウィンドウに戻り、起動直後では白紙の編集画面に変わります。

＜一致条件＞

検索対象は単語ですが、前方部分一致条件、後方部分一致条件での照合もアステリスクを使用することにより可能です：

例えば、con の綴りで始まる全ての語（con 自身を含めて）をサーチするには con* をキーワードに指定します。同様に語尾 ito で終わる語は *ito を検索語とすることで発見することができます。また、*ito* のように両条件を組み合わせると語中、語頭、語尾を問わない文字列の照合として利用できます。

テキストの文字列と検索キーワードは次の場合に一致したと見なされ、その文字列を含む文が出力されます：

1．テキスト文の文頭での大文字はキーワードの大小文字の両方にマッチする

2．テキスト文内では大文字・小文字の区別が有効

この結果、キーワードを一部または全部大文字で指定した場合はテキスト側で全く同様な大小文字を使用していないと一致したとは見なされません。例えば、テキストの単語 EJEMPLO はキーワードの EJEMPLO とのみ一致し、一方、テキスト文頭の Ejemplo, ejemplo はキーワードの ejemplo で検出することができます。

＜検索結果ウィンドウ＞

検索結果を出力する場面では上部のタイトルバー内に検索した「単語」が表示されます。該当語が検出されないときは、出力ウィンドウにWORD NOT FOUND! と指示します。検索語が発見されれば、各事例毎にそのページ、頁頭からの行数を添えて、その語を含む「文」を表示します。結果テキストはウィンドウの大きさに応じて自動的に word wrap します。1画面に収まらないときは、右のスクロールバーで上に巻き上げて見ることが出来ます。

結果データは直ちに削除・挿入を加えるなど自由に編集することが可能です。DOS-V 機の Windows で使用している場合は、ここで全ての欧州特殊文字を入力表示することが可能です。機種を問わずスペイン語特殊文字は次の方法で dead key 入力でき、スペイン語 font が表示されます。

アクセント付き文字ははじめに@を押し次に母音を打鍵する
ñ, Ñは　　はじめに@を押し次にn, Nを押す
逆疑問符、感嘆符は　はじめに@を押し次に？，！を押す
ü, Üは　　はじめに ¨ を押し次にu, Uを押す

またマウスのドラッグによる領域指定でクリップボードを介してCut & Paste も行えます。上部のメニューバーをクリックして次のいずれかのステップを選択して下さい：

[Word Search]　新たな検索をする。現在の検索結果データは失われます
[Save File]　現在の検索結果をファイルに保存する
[Quit]　プログラムを終了する
[About]　プログラムのバージョンと日付を表示します
[Help]　使用法の説明文書を表示します

◆◆その他の仕様◆◆

検索結果データは最大26Kbまでです。検索対象ファイルの大きさには制限がありませんが、高頻度の語を検索した場合、出力量がオーバーする可能性があり、その時は検索を打ち切り、出力データの最後部に“Word Search Interrupted”を書き込みます。

コンテキスト文の長さは最大1024文字（空白も含む）で、それ以上の場合は後部をカットします。発見位置を示す行数は概略で、その文の末尾の位置の目安となります。

編集画面のフォントを変更することはできません。

＜動作確認＞

IBM 互換機で、IBM DOS 5.0J／V 上の IBM 版 Win 3.0及び Microsoft 版Win 3.1で走行を確認しました。

◆◆プログラム概要◆◆

プログラム名　wwfind
ファイル名:　wwf.exe
Version:　0.9
種類：　テキスト文字列検索プログラム
動作環境：　日本語 MS-Windows ver 3.0 以上
作者：　出口厚実
ファイルサイズ：21565 bytes
ファイル日付：　1994.03.10
開発言語：　Turbo C++ for Windows ver. 3.1

◆◆改版履歴◆◆

ver 0.7(1993.07.12)　公開初版

0.8(1993.08.12)　いくつかのバグ取りのほか、次の諸点を改良した：安全チェックポイントを増やした；ページ付けの無いテキストを考慮して、ページヘッダの初期値を '---' にした；ページ標識を誤認した場合、30字で打ち切るように変更；オーバーフローのために検索を中断したときその旨指示を出すようにした；

0.9(1994.02.20)　save 時の検索語表示を削除した；検索結果出力をワードラップするように変更した；

0.91(1994.03.10)　この document をヘルプメニューで表示できるようにした

◆◆免責◆◆

作者はこのプログラムの仕様内容及びその信頼性を保証するものではありません。不具合いやその他お気づきの点・御要望は原作者までご連絡下されば、次期バージョンで解決するよう努力致しますが、乏しい力量のゆえにご期待にそえないこともあります。

◆◆再配布◆◆

"研究教育機関に所属する個人またはグループが学術研究の目的で作成したデータ及びプログラム（コード）を広く無償で（金銭・物的等価要求を伴わずに）公開し合い相互利用し合う"ことに賛同される方々の間では、この趣旨に従って本プログラムを自由に使用し、再配布出来ます。ただし、再配布されるときは、この document file を含めて一組でなされるものとし、そのことを原作者にご連絡下さるようお願いします。

1994. 03. 10　出 口 厚 実

3．2　WWFI.EXE（Win WordFinder i)『英語版MS-Windows対応欧文単語検索プログラム』

wwf.exe がかかえる問題点を取りあえず、海外版 Windows 上で解決できないものかと考えて、試作したのが姉妹ソフト wwfi.exe である。ただし、対象とするテキストの文字コードは wwf.exe と異なるため、文字列照合ルーチンも別になっていて、両者の使用目的を明確に区別して用いないと検索漏れが生じる恐れもあるので注意を要する。プログラム自体は相反する土俵で、すなわち wwf.exe を英語 Win で、wwfi.exe を日本語 Win で動作させることも可能である（この場合でも特に警告を出さない）が、あべこべに用いるとそれぞれに期待されている機能を引き出すことができないケースが起こる。

この検索ツールの特徴は、Windows 用ソフトでありながら“非ウィンドウズの DOS 標準テキスト”専用に設計した点にある。これまでにIBM 互換機で用いられる一般的なテキスト（コードページ437）は英語 Windows の画面では、そのままでは文字化けをして正常に表示されない。そこで、海外版ワープロやエディタは画面フォントの切り替え（IBM OEM フォントに設定し直す）をサポートしているのが普通である。現在、Windows のansiコードで作られたテキストが乏しい状況を考え、デフォルトでいきなり OEM 形式テキストを走査するような検索ソフトも必要ではないかと思ったのである。

そのために、入力された検索語を Windows 関数の AnsiToOem を呼び出して OEM に変換した後でテキスト文と照合する手順を踏んでいる。英語版 Windows 上では、当然のことであるが、スペイン語を含め欧文特殊文字の入力はダイアログボックスでも可能で、従ってテキスト文と同等の感覚で使用できる。

その他の操作性、機能、出力最大量の制限 etc. は前項の wwf.exe に準じるので省略するが、stx テキストと同様なページ標識の（すなわち［ ］を用いる）wwfi.exe の他にも、OCP (Oxford Concordance Program) フォーマットの、文書名、ページ標識をもつ文書に対応した別バージョン wwfit.exe も用意している。

3．3　WWSCH.EXE（Win WordSearch）『日本語 Windows 対応：欧文検索エディタ』

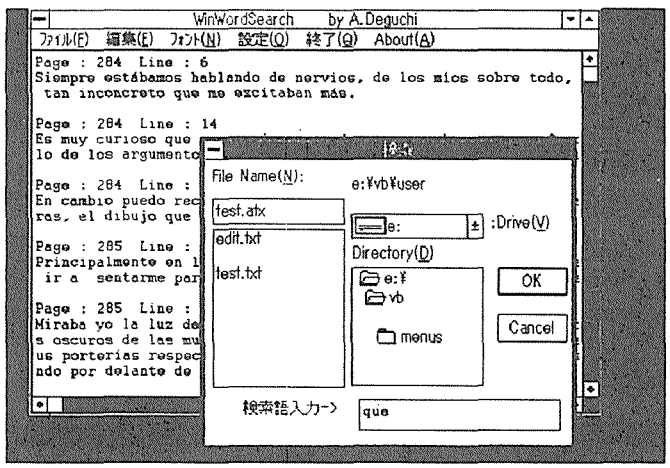
上記の２つの自家製ソフトはＣ言語（Turbo C＋＋ for Windows ver. 3.1）を利用したものであるが、Visual Basic for Windows でどの程度実用的なプログラムが書けるのかと、いう好奇心も手伝って、試験的に製作したのがこのソフトである。結論の方を先に述べれば、小規模なテキストを対象とするならばともかく、大きなファイルを常時検索するには適さない。少なくとも、wwf.exe, wwfi.exe に比較するとファイル検索の照合に要する時間が格段に長くなる。検索エンジンの部分も最初は Visual Basic でコードを書いてみたが、文の切り出しとマッチングで絶望的にスピードが低下するので、パタン照合関数は別に Turbo C＋＋ で書いたDLL を呼び出すように改めたのが、現在の version である。

起動直後のウィンドウでテキストの新規編集が可能であるが、メニューバーから「ファイル」

「編集」「フォント」「設定」「終了」「About」をクリックして選択できる。「ファイル」のサブメニューとして［検索］［オープン］［保存］［印刷］［終了］のいずれかが処理できる。「検索」、「オープン」、「保存」では wwf.exe などと同様にマウス操作によってドライブ・ディレクトリ・ファイル名を指定する。

ファイル検索を何度も繰り返していると、読み込みのためにオープンしたはずのファイル名ボックスをうっかりと保存用の結果出力ファイル名として勘違いして承認ボタンを押してしまい、原テキストファイルを消してしまうというミスを犯すことがあるので、「検索」・「オープン」・「保存」の各場面で、ダイアログボックスの背景色を区別することで差を意識させるという工夫をした。このようなアトリビュートの変更が容易なのはVBプログラミングの利点の１つに違いないだろう。

［検索］はオンメモリのテキストに対してではなく、新規に外部テキストを読込で行ない、結果をウィンドウに出力する。なお、wwsch.exe では検索語入力ボックスにおいても、全ての欧文特殊字が使用可能である。また、スペイン語に関しては下記（４）の方法で dead key 入力も出来るようにしてある。



［図２］wwsch. exe で検索ダイアログを表示しているところ

「編集」のサブメニューを選ぶとマウス操作により cut, paste, copy ができる。

「フォント」の下位メニューから、更に［フォント種類］［フォントサイズ］に分かれ、前者では、Courier, Roman, Arial, Helvetica の４種を選択使用できるようにした。またフォントサイズでは12, 24ポイントのいずれかが利用可能である。

もう１つのユーザ設定部分は、メニュー［設定］から選ぶことのできる背景色である。デフォルトは白であるが、青または灰色に変更可能なようにしてある。

テキストの編集において ansi フォントの中にある非英欧文特殊字はすべて表示可能である。ただし、NEC9800 シリーズなど非 DOS／V 上の日本語 Windows では、表示はされるが、実

際にはキーボードから入力できないという制約がある。スペイン語に関しては独自のキー入力監視とdeak keyを支援する回路を設けているので、ＮＥＣを含む全ての Windows システムで特殊字が簡単に入力できるはずである。この場合の、キーの割り当ては以下のようにしたので、wwf.exe, wwfi.exe と一部異なる；

（4）

アクセント付き小文字、大文字母音	“;” を先行打鍵し次に母音を入力
ñ, Ñ	“;” を先行打鍵し次にn, Nを入力
逆疑問符	@
逆感嘆符	|

WWSCH には検索機能も設けておいたが、むしろ、いくつかのフォント種類、背景色の切り替えなど、日本語 Windows の『メモ帳』にない機能を使って、スペイン語対応の短いテキストや備忘録作りに使用できる簡易エディタとして役立つかもしれない。簡易エディタとして見た場合、edit controlのおかげで、他の Windows ソフトとの間で cut & paste も可能であり、スクロールも比較的スムーズで特に不便を感じない。

3．4　DFIND.EXE（D & D Find）

『日本語 Windows 対応 Drag & Drop 欧文テキスト検索』

wwf.exe／wwfi.exe／wwsch.exe の３プログラムにより、日本語 Windows の中で欧文テキストを検索したり、その結果を加工したりする作業がかなり改善された。しかし、巨大なファイルに対するヒット数の多い検索出力がオーバーフローする制限事項はやはり課題であった。DOS上の検索でも、無論、同様な問題が生じ得るが、MS-DOS では高速、高機能な File browser やエディタが存在するため、検索ツールをそれらに直結する便法で、能率よい環境を構築できる。例えば、拙作 wfind.exe 2.0 では予め環境変数にセットしたファイル閲覧ソフトあるいはeditorを検索ソフト内部から呼び出し、一度ファイルに書き出した出力を自動的に開いて起動するような仕様にした。この手法はほとんどスピードの低下を招くことなく結果画面をみたり、編集に移行できるという利点がある。当然、検索結果出力はほぼ無制限といってよい程（i. e. ディスクスペースが許す限り）の余裕がある。

Windows 用で、かつ非英欧文の大ファイルに対応できる editor は数える程しかなく、筆者自身の手に馴染むものを未だ見出せないでいる。いくつかのものはフォントの切り替えでスペイン語等を表示できる能力をもつが、起動後でしかフォント変更が効かず、また１度終了すると日本語モードに戻ってしまうというものもある。検索プログラムの内部から呼び出すにはどうしても立ち上げ時から指定フォントに設定可能なものでなければならない。更に起動時間が最短でなければならないという要件も重要である。

もう1つ改良すべき点は複数ファイルの一括検索である。DOS 版 wfind 2.0 では検索対象ファイルを複数指定したり、ファイル名にワイルドカードを用いることもできる。複数ファイルのサポートは検索ファイルの選択画面での処理を拡張する方法もあるが、より GUI にふさわしいドラッグ&ドロップ処理にできないものかと考えた。すなわち、ドライブ、パス、ファイル名の指定作業はファイルマネージャに任せ、選択表示されている1つ以上のテキストファイルを検索プログラムの上にドラッグして重ねることによって検索を自動的に開始させるという方式である。

以上のようなコンセプトの実現を目指したのが dfind.exe で、もし、今後、実用的な高機能 editor が多数出現して、選択の幅が広がるようになれば、さらに利用価値が増すのではないかと思われる。プログラムの中心は実質的に検索エンジンとファイル読み込み・書き出し処理に限定されていると言ってもよく、コーディングの面倒なファイル情報の入手、及び結果の閲覧・編集を他のアプリケーションに委ねるという、スリムな設計が可能になる。その仕様、機能の概略は以下の通りである：

（5）

起動方法：　他の一般ウィンドウズ用アプリケーションと同様

検索方法：　1．検索語をボックスに入力する

2．起動前、または起動後に開かれたファイルマネージャのファイル・リストの中からファイルをドラッグして、D&D FIND のウィンドウ領域にドロップする。複数のファイルを検索対象にしたい場合は、隣のファイルであれば Shift キーを押し下げて、そうでないときは Ctrl キーを押し下げてマウスをクリックすればファイル名が選択状態になるので、その後1度だけ、drag & drop すればよい。

3．同じ検索語で再検索をする場合は、ドラッグ&ドロップを繰り返す

4．同じファイル（群）に別の語を検索し直すには、検索語を再入力してから検索ボタンをクリックする

検索語入力：欧文特殊文字の表示入力が可能；スペイン語に対しては dead key による入力をサポート；キー割り当ては wwf.exe と同じ

検索出力：　1．該当例が発見されると、直ちに指定 editor を呼び出し、テキストファイル名、該当語例を含む全文を、そのページ、行数と共に表示

2．テキストは80桁に強制改行

3．複数ファイルを指定して検索した場合、事例が発見されなかったファイルはそのパス・ファイル名のみ表示される

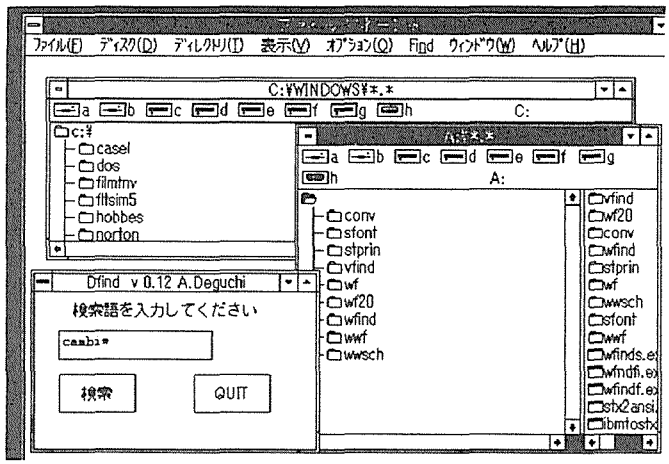
4．出力総量の上限はディスク上の空きスペースに依存

5．指定 editor の文書容量による制限を受ける；ただし editor 側が決める容量をオーバーしてもテンポラリーファイルとしての結果は残る

6．指定 editor が MDI を採用していれば、検索ごとの結果をクローズせずに

ウィンドウを多重に開いておくことが可能

終了：　　「終了」ボタンまたはプログラムウィンドウの「閉じる」をクリックする



[図3] ファイルマネージャと dfind.exe を並べて検索ファイルをドロップする

4．データ互換性のために

テキストデータを処理するための土台として Windows がそれ程急速に普及しない理由として、関連ソフトの不足の他にも、Windows 形式のデータへの対応という足枷がある一面も見逃せない。本節では非ウィンドウ用テキストファイルを Windows 下でも有効利用するための対策について述べる。

4．1 Windows 用テキストデータ

テキストデータの形式は効率よくまた正確にテキストを処理するための基本要素でもある。機種や OS の違いに左右されない一貫したフォーマットを使いたいという要望があるのはもっともなことである。不必要な混乱や誤検索を避けるためにも、この可搬性を優先するポリシーはなるべく尊重されるべきである。しかし、外国語文データに関しては、止むを得ないケースを除いて、その言語で用いられる正書法イメージから離脱したコード体系は、例え可搬性の一部が損なわれようとも採用されるべきではないというのが筆者の考え方である。

その表示が物理的に不可能でないのに、アクセント付き文字を“他のシステムとの互換性維持のために”わざわざ、a／とか a' や 'a または @a と表記するのは、テキストの可読性を著しく損ない、また入力・校正ミスを誘うのみならず、テキストに対する利用者の心理的抵抗や疎遠感を増幅する。

Windows においては、DOS で困難であった欧文特殊字のコード化が曲がりなりにも実現さ

れているのであるから、その利点を傍観する理由はない。Windows でのテキスト処理のために、データ変換という余分な過程が必要になるとしても、画面に正書法通りの文字を読めるという快適さのほうが大切だと思われる。非Windows からのデータ変換の実作業は、例えば、出口(1993) で紹介した方法でさしたる困難を伴わずに実現できる。一時的にウィンドウ環境をテストしてみる場合だけでなく、他のプラットホームから Windows へ全面移行するケースでも、よほど膨大な量のテキストを扱わない限り変換操作はごく短時間に終了するであろう。

4．2 DOS 等のテキストとのデータ互換

スペイン語データに限っても、現在、多種類のDOS テキストが利用され、流通している。データ形式は各人のハード環境に応じて適切な方式を選んだり工夫すればいいが、前稿（1993）の汎用コンバータtxcnv.exe, tx2cnv.exe を利用すれば、自分がよく用いるタイプの変換専用の実行プログラムが簡単に作れるので、これをDOS 側に常備しておけばテキストコンバート作業はほとんど手間がかからない。変換方式で大別すると、複文字への換字を行うものと、コード番号の入れ替えのみで済ませられる場合がある。上記プログラムをカスタマイズして机辺で必要なテキスト変換のために、以下のような Windows 用テキストとの双方向変換実行プログラムを作ってみた。

（6）

ibm2ansi.exe	コードページ 437 → Windows 形式
ansi2ibm.exe	Windows 形式 → コードページ437
sibm2ans.exe	コードページ 850 → Windows 形式
ans2sibm.exe	Windows 形式 → コードページ 850
stx2ansi.exe	stx テキスト → Windows 形式
ansi2stx.exe	Windows 形式 →stx テキスト

他のOS用テキストの場合はテキストフォーマットやメディア規格の違いも加わって、文字コードの変換のみでは互換を取れないのが普通であるが、フロッピーディスクなどの外部記録媒体に対する何らかの変換ユーティリティがシステムに付属しているケースが多い。例えば、Macintosh で作られたテキストデータであっても、もしそれがstx 形式に則っていればDOS 形式にファイル変換された後に、上のstx2ansi.exe で直ちに Windows 用のデータに変換して利用することが可能である。また、欧文特殊字を含んでいるマッキントッシュのplain テキストに対しては、やはり拙作のsettcn.exe を利用して専用のMAC／Windows 間のテキストコンバータ（Cf. （7））を作出してあるが、これらも常備しておくと便利であろう。

（7）

mac2ansi.exe	Macintosh テキスト → Windows 形式

ansi2mac.exe　　Windows 形式 → Macintosh テキスト

5．結語

普及型のパソコンで利用できる代表的な GUI 環境である Windows において、外国語テキストを検索して言語研究や教育の素材として利用する際に生じるいくつかの問題点を取り上げ、その解決を目指す二、三の具体的方法を検討し紹介した。ハードウエアの目覚ましい進化に比べると、この分野では、“ユーザにやさしい” ソフトの開発の歩みはまだ遅い。GUI 自体は特にテキスト処理と馴染みにくい性質を持つものではなく、むしろ、今後、その特長を生かした種々のユーティリティが作成され、一般利用者がDOS 環境よりも容易にテキストデータを扱えるようになることを期待したい。　　(1994.5.1)

[注]

1．存在するデータの総量が乏しいかどうかは対象言語によって差があるのは当然だが、さまざまな社会的要因によって、言語研究者が互いに利用できるテキスト・データの質と絶対量が限られていて、急速な改善が見込まれないという現状に甘んじなければならない。幸い、スペイン語学に関しては国内で、このような目的のためのデータプールの機関が発足し、これからデータ処理を始めようとする人々にも広く門戸を開放されている。

2．拙稿（1990ab, 1991, 1993, 1994）ではMS-DOSの環境下で外国語データ検索を容易にするための方策を論じ、具体的に自作ソフトの開発例を紹介した。

3．ASCII 誌（Vol.17 No.6, pp.406-9）でも確認されているので、どの機種の日本語 MS-Windowsにも見られる現象であろう。なお、同誌 Vol.17. No.8 によれば、英語版 Word 2.0 で作成したデータを日本語版 Word 5.0 に読み込めば日本語 Windows でスペイン語が正常に表示できるという（p.444）。

4．例えば、次のような検索ソフトが存在する：

wgrep.exe／GREP の正規表現が使えるが、コンテキストとして得られるのは 1 行と行番号のみである。外部viewerを指定でき、結果はマルチファイルとしてタイルまたはカスケード表示のどちらかが選べる。検索スピードは遅い。

wingrep.exe／1 つのウィンドウに連続して結果を出力する。複数ファイルに対応するが文脈は得られず、また行番号ではなくファイル頭からの16進値offsetが示されるだけなので実用的でない。Drag & Drop 方式だが、検索速度は非常に遅い。

find.dll／ファイルマネージャへのアドオンソフトで、Directory 全体など多数のファイルを 1 度に検索するのに向いている。高速だが出力はファイル名、行番号と 1 行文脈のみで、検索対象も固定文字列に限られる。

5．このソフトは他の拙作の DOS 版検索ツールと共に、東京スペイン語データバンクに登録済みである。

＜REFERENCES＞

出口厚実　(1989) スペイン語テキストデータとパーソナルコンピュータの使用環境-Estudios Hispánicos 14, pp.1-13
－　(1990a) スペイン語テキストファイルの作成とファイル形式の変換-Estudios Hispánicos 15, pp.1-15
－　(1990b) スペイン語テキスト処理の実際：単語検索の諸問題－大阪外国語大学論集 4, pp.137-152
－　(1991) スペイン語動詞屈折形態の同定と探索-Estudios Hispánicos 16, pp.15-28
－　(1993) テキストデータの形式とその変換：スペイン語の場合-Estudios Hispánicos 18, pp.45-59
－　(1994) 外国語テキスト処理に関するユーティリティの自作と活用－大阪外国語大学での情報処理・研究のあり方について, pp.1-8

(1994年 5 月 9 日　受理)